Pattern Recognition 43 (2010) 3674–3682

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Quasi-supervised learning for biomedical data analysis

Bilge Karaçalı *

Electrical and Electronics Engineering Department, Izmir Institute of Technology Urla, Izmir 35430, Turkey

ARTICLE INFO

Article history:
Received 3 September 2009
Received in revised form
30 March 2010
Accepted 29 April 2010

Keywords:
Biomedical data analysis
Abnormality detection
Nearest neighbor rule
Support vector machines
Flow cytometry
Electroencephalography

ABSTRACT

We present a novel formulation for pattern recognition in biomedical data. We adopt a binary recognition scenario where a control dataset contains samples of one class only, while a mixed dataset contains an unlabeled collection of samples from both classes. The mixed dataset samples that belong to the second class are identified by estimating posterior probabilities of samples for being in the control or the mixed datasets. Experiments on synthetic data established a better detection performance against possible alternatives. The fitness of the method in biomedical data analysis was further demonstrated on real multi-color flow cytometry and multi-channel electroencephalography data.

1. Introduction

Pattern classification methods constitute the backbone of biomedical data analysis on high dimensional quantitative data provided by the state-of-the-art medical imaging and high-throughput biology technologies. The general strategy relies on expert-curated ground truth datasets providing the categorical associations of all available data samples. Ground truth datasets are then used as the basis for statistical learning, specifically to construct a classification rule using one of a host of methods such as the support vector machines [1,2], nearest neighbor classifiers [3,4], neural networks [5], discriminant functions [6] and so on.

A host of problems exist in obtaining manually labeled ground truth datasets on biomedical data for supervised learning. The most pertinent problem is the cost associated with the task: Manual processing of biomedical data is often very laborious and requires long hours of dedicated expert effort. Additional deterrent factors are the systematic bias incorporated into the learning system in emulating the decision criteria of a fixed expert group, the deficiency of manual processing of multivariate data, and the uncertain characterization of the learning problem by the often modest-sized ground truth data.

These considerations necessitate a novel formulation for statistical learning on biomedical data requiring no more than elemental and easily obtainable expert interaction. In the literature, quasi-supervised learning refers to learning strategies that deal with prominently unlabeled data, where some labels are available and only through indirect user interaction [7,8]. Processing of unlabeled data in classification tasks is studied by semi-supervised learning, where the principal aim is to obtain better characterization of the posterior class distributions by taking unlabeled data into account [9].

In this paper, we address a target identification problem where labeled samples are available from one class only, in a control dataset. A second, unlabeled dataset is also provided and contains a mixture of samples from both control and target classes. Identification is carried out using a computational algorithm that contrasts the unlabeled mixed dataset samples to the control dataset, and selecting those that are dissimilar from the control samples beyond a statistical significance level as target samples. Note that the lack of class labels other than the control dataset place this problem beyond the premises covered by semi-supervised learning and more towards unsupervised learning. Note also that as such, this strategy accommodates the biomedical data analysis task well: Often, a dataset of control samples is very easy to obtain while representative abnormalities require manual identification. In computational analysis of histology slides, for instance, a pathologist can easily identify tissue cross-sections that are free from cancerous abnormalities. Such abnormalities, on the other hand, occur amid tissue that is benign in appearance, and have to be either painstakingly labeled by an expert pathologist from whole slides in a computerized system or imaged selectively from within tumor boundaries.

The next section is devoted to the derivation of an algorithm that contrasts two datasets by computing estimates of the

* Tel.: +90 232 750 6719; fax: +90 232 750 6599.
E-mail address: bilgekaracali@iyte.edu.tr
URL: http://www.iyte.edu.tr/~bilgekaracali

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.04.024

posterior probabilities of samples belonging to one or the other. Section 3 presents a quantitative performance evaluation of the algorithm in comparison to alternative strategies from graph theory and support vector machine classification on synthetic data. Performance evaluation experiments are followed by the application of the proposed quasi-supervised learning method to the analysis of real multicolor flow cytometry data for comparing cell distributions, and to the identification of brain activity patterns associated with different visual stimuli in real multi-channel electroencephalography data.

2. Methodology

In this section, we construct a numerical algorithm that realizes a quasi-supervised learning strategy using the asymptotic properties of a nearest neighbor classification rule. These properties lead to the derivation of nonparametric estimates for the posterior probabilities and subsequently of several measures for class overlap on the basis of each sample.

2.1. Likelihood ratio estimation via the nearest neighbor rule

Given a reference set $R=\{x_i,y_i\}$ of points $x_i \in \mathcal{X}$ and their respective class labels $y_i \in \{0,1\}$ for $i=1,2,\ldots,\ell$, a nearest neighbor classifier is defined by

$$F_R(x)=y_{i_0} \quad \text{with } i_0=\arg\min_{i=1,2,\ldots,\ell} d(x,x_i) \tag{1}$$

for $x \in \mathcal{X}$, where $d(\cdot,\cdot)$ denotes the distance metric on $\mathcal{X}$ [3]. The nearest neighbor classifier in Eq. (1) has been a benchmark classification method in the pattern recognition literature thanks to its simplicity and to several asymptotic properties linking its error rate to that of the optimal Bayes rate. Indeed, it can be shown that the asymptotic error rate of the nearest neighbor classifier is bounded from above by twice the Bayes rate [10]. Here, we will consider the ratio of the classification decisions for a point x during the course of successive classifications each time using a different reference set, as the number of classifications grows large. Below, we argue that the fraction of times a given point x is assigned to a given class provides an estimate of the posterior probability associated with that class, and derive an algorithm to compute it analytically without carrying out all possible nearest neighbor classifications.

2.1.1. Exhaustive nearest neighbor classification using random reference sets

Let $\{R_j\}$, $j=1,2,\ldots,M$, be a collection of independent and identically distributed reference sets, consisting of n points from each of the two classes. Define $f_0(x)$ and $f_1(x)$ by

$$f_0(x)=\frac{\sum_{j=1}^{M}\mathbf{1}(F_{R_j}(x)=0)}{M} \tag{2}$$

and

$$f_1(x)=\frac{\sum_{j=1}^{M}\mathbf{1}(F_{R_j}(x)=1)}{M} \tag{3}$$

where $\mathbf{1}$(statement) is 1 if statement holds and 0 otherwise. The critical observation is that for sufficiently large M,

$$f_0(x)\simeq\frac{p(x|x\in\mathcal{C}_0)}{p(x|x\in\mathcal{C}_0)+p(x|x\in\mathcal{C}_1)} \tag{4}$$

and

$$f_1(x)\simeq\frac{p(x|x\in\mathcal{C}_1)}{p(x|x\in\mathcal{C}_0)+p(x|x\in\mathcal{C}_1)} \tag{5}$$

providing

$$\frac{f_0(x)}{f_1(x)}\simeq\frac{p(x|x\in\mathcal{C}_0)}{p(x|x\in\mathcal{C}_1)} \tag{6}$$

where $p(x|x\in\mathcal{C}_0)$ and $p(x|x\in\mathcal{C}_1)$ represent the class conditional probability densities for the classes $\mathcal{C}_0$ and $\mathcal{C}_1$, respectively [4,10]. The basis of this observation is as follows: Let $\mathcal{N}(x)\subset\mathcal{X}$ be a small spherical neighborhood of x of size $V(\mathcal{N}(x))$, and R be a random reference set containing n points from each of $\mathcal{C}_0$ and $\mathcal{C}_1$. Then, $\mathcal{N}(x)$ will contain approximately $n\cdot V(\mathcal{N}(x))\cdot p(x|x\in\mathcal{C}_0)$ points of R from $\mathcal{C}_0$ and $n\cdot V(\mathcal{N}(x))\cdot p(x|x\in\mathcal{C}_1)$ points from $\mathcal{C}_1$. In addition, since these points are the closest in R to x, nearest neighbor classification of x with respect to R reduces to one that uses these $n\cdot V(\mathcal{N}(x))(p(x|x\in\mathcal{C}_0)+p(x|x\in\mathcal{C}_1))$ points as the reference set. Furthermore, since $p(x|x\in\mathcal{C}_0)$ and $p(x|x\in\mathcal{C}_1)$ are both approximately constant over $\mathcal{N}(x)$, these points can also be assumed to be uniformly distributed in $\mathcal{N}(x)$, partitioning $\mathcal{N}(x)$ into distinct $\mathcal{C}_0$ and $\mathcal{C}_1$ regions with volumes

$$\frac{V(\mathcal{N}(x))\cdot p(x|x\in\mathcal{C}_0)}{p(x|x\in\mathcal{C}_0)+p(x|x\in\mathcal{C}_1)}$$

and

$$\frac{V(\mathcal{N}(x))\cdot p(x|x\in\mathcal{C}_1)}{p(x|x\in\mathcal{C}_0)+p(x|x\in\mathcal{C}_1)}$$

respectively. Finally, as the probability of x being in one or the other region is proportional to their volumes, Eqs. (4) and (5) follow. This idea is illustrated in Fig. 1. Note also that since the inclusion of an equal number of samples from each class in the reference set leads to equal prior probabilities for $\mathcal{C}_0$ and $\mathcal{C}_1$, in effect, $f_0(x)$ and $f_1(x)$ in Eqs. (4) and (5) compute estimates of the posterior distributions of the classes $\mathcal{C}_0$ and $\mathcal{C}_1$ at the point x. Under these circumstances, their ratio coincides with the likelihood ratio of the class conditional probabilities as well.

This observation suggests that given a point $x\in\mathcal{X}$, the likelihood ratio as well as the posterior probabilities of the two

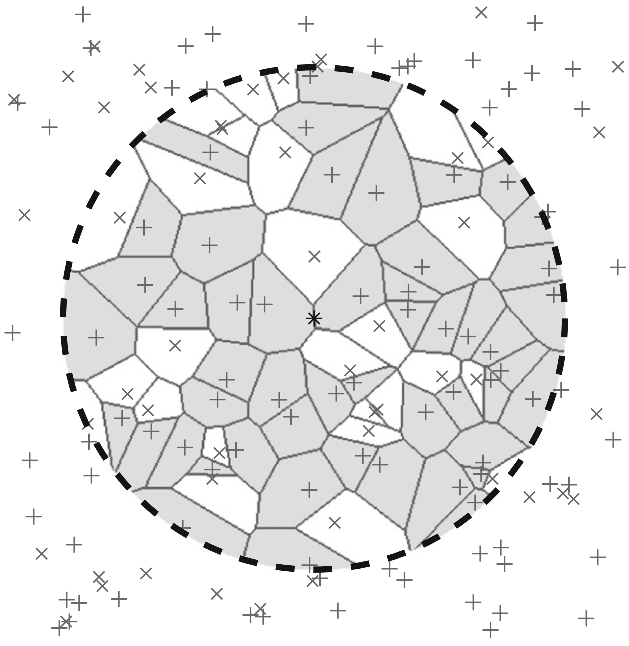

Fig. 1. Illustration of the asymptotic property of the nearest neighbor classifier explored by the quasi-supervised learning strategy. Around a point x represented by $\star$, $p(x|x\in\mathcal{C}_1)=2p(x|x\in\mathcal{C}_0)$, leading to nearly twice as many $\mathcal{C}_1$ points (49) as $\mathcal{C}_0$ points (26), represented by + and ×, respectively. As a result, the $\mathcal{C}_1$ partition in the neighborhood of x leads the $\mathcal{C}_0$ partition in area 0.6721 to 0.3279, making x twice as likely to be assigned to $\mathcal{C}_1$ as to $\mathcal{C}_0$.

classes at x can be estimated based on a dataset $\{x_i,y_i\}$ by carrying out multiple nearest neighbor classifications on x using randomly chosen reference sets from $\{x_i\}$ with equal representation from both classes, and keeping track of the number of times x is assigned to $\mathcal{C}_0$ and to $\mathcal{C}_1$. Such an approach has previously been used to identify cancer-related abnormalities in a set of Hematoxylin and Eosin stained breast cancer histology slides by carrying out repeated nearest neighbor classifications followed by regression on the estimated log-likelihood ratio to reduce the estimation noise [11]. The issue of determining how many such random nearest neighbor classifications would be required for a given recognition problem, however, was not addressed, nor was the computational expense related to carrying out a large number nearest neighbor classifications. These issues are addressed here as follows.

The reliability of the above estimate of the log-likelihood ratio at a point x clearly depends on M, the number of successive random nearest neighbor classifications. With n points from each class in the reference set, the total number of distinct reference sets of size $2n$ is given by

$$\binom{\ell_0}{n}\binom{\ell_1}{n}$$

where ℓ_0 and ℓ_1 denote the number of points in the set $\{x_i\}$ belonging to the respective classes. This number is maximized for

$$n=\left\lfloor\frac{\ell_0\ell_1+\ell_0+\ell_1+1}{\ell_0+\ell_1+2}\right\rfloor$$

Even for modest reference set sizes, say of 100 samples each, the number of distinct nearest neighbor classifications that can be carried out reaches levels over 10^{50}, well beyond the abilities of the present day computing equipment.

2.1.2. Analytical computation of the posterior probabilities

While carrying out an exhaustive evaluation of all possible random nearest neighbor classifications is not feasible, it is still possible to compute the average number of times a given point would be assigned to either class at the end of such an evaluation. Consider the distances $d_i{=}d(x,x_i)$ between a given point x and each x_i for $i=1,2,\ldots,\ell$. Let $d_{(i)}$ denote the ordered sequence of all $\{d_i\}$ with $d_{(1)}\le d_{(2)}\le\cdots\le d_{(\ell)}$, and $\{x_{(i)}\}$ and $\{y_{(i)}\}$ be such that $d_{(i)}{=}d(x,x_{(i)})$ and $y_{(i)}$ is the class label of $x_{(i)}$. After an exhaustive nearest neighbor analysis, $f_0(x)$ represents the probability $\Pr\{y{=}0\}$ of assigning x to the class $\mathcal{C}_0$ based on a reference set R with n points from both classes selected randomly from the $\{x_i\}$,

$$f_0(x)=\Pr\{y=0\} \tag{7}$$

This probability can be decomposed conditionally on whether or not the point $x_{(1)}$ is in R, providing

$$f_0(x)=\Pr\{x_{(1)}\in R\}\mathbf{1}(y_{(1)}=0)+\Pr\{x_{(1)}\notin R\}\Pr\{y=0|x_{(1)}\notin R\} \tag{8}$$

since $\Pr\{y=0|x_{(1)}\in R\}$ is 1 if $y_{(1)}{=}0$, and 0 otherwise. For notational simplicity, define E_k as the joint event $x_{(1)},x_{(2)},\ldots,x_{(k)}\notin R$. Carrying the same decomposition strategy further to $\Pr\{y=0|E_1\}$ provides

$$\Pr\{y=0|E_1\}=\Pr\{x_{(2)}\in R|E_1\}\mathbf{1}(y_{(2)}=0)+\Pr\{x_{(2)}\notin R|E_1\}\Pr\{y=0|E_2\}$$

In general, therefore, for $\Pr\{y=0|E_{k-1}\}$,

$$\Pr\{y=0|E_{k-1}\}\quad=\Pr\{x_{(k)}\in R|E_{k-1}\}\mathbf{1}(y_{(k)}=0)+\Pr\{x_{(k)}\notin R|E_{k-1}\}\Pr\{y=0|E_k\} \tag{9}$$

Putting the whole sequence of conditional probability decompositions into the original equation for $f_0(x)$, we obtain

$$\begin{aligned}f_0(x)=&\Pr\{x_{(1)}\in R\}\mathbf{1}(y_{(1)}=0)+\Pr\{x_{(1)}\notin R\}(\Pr\{x_{(2)}\in R|E_1\}\mathbf{1}(y_{(2)}=0)\\&+\cdots+\Pr\{x_{(\ell-1)}\notin R\}(\Pr\{x_{(\ell)}\in R|E_{\ell-1}\}\mathbf{1}(y_{(\ell)}=0)\\&+\Pr\{x_{(\ell)}\notin R|E_{\ell-1}\}\Pr\{y=0|E_\ell\})\ldots)\end{aligned} \tag{10}$$

Note that $\Pr\{E_\ell\}=0$ since the reference set R must have at least $2n$ data points. In fact, the decomposition need not be carried out beyond some $k^\star$ given by

$$k^\star=\max\left\{k\left|\sum_{k'=k}^{\ell}\mathbf{1}(y_{(k')}=0)\ge n \text{ and } \sum_{k'=k}^{\ell}\mathbf{1}(y_{(k')}=1)\ge n\right.\right\} \tag{11}$$

since $\Pr\{x_{(k^\star)}\in R|E_{k^\star-1}\}=1$ and $\Pr\{x_{(k^\star)}\notin R|E_{k^\star-1}\}=0$. The algebraic development above can be repeated for $f_1(x)$ in an identical manner.

This derivation suggests the following algorithm to compute $f_L(x)$ for $L\in\{0,1\}$ for a given x based on the dataset $\{x_i,y_i\}$, $i=1,2,\ldots,\ell$ and a fixed n:

- Compute $d_i{=}d(x,x_i)$.
- Sort $\{d_i\}$ so that $d\ (1)\ \le d_{(2)}\le\cdots\le d_{(\ell)}$, and determine the corresponding sequences $\{x_{(i)}\}$ and $\{y_{(i)}\}$.
- Identify $k^\star$ in Eq. (11), and set $\Pr\{y=L|E_{k^\star-1}\}=\mathbf{1}(y_{(k^\star)}=L)$
- For $k=k^\star-1,k^\star-2,\ldots,1$, compute $\Pr\{y=L|E_k\}\ =\Pr\{x_{(k+1)}\in R|E_k\}\mathbf{1}(y_{(k+1)}=L)+\Pr\{x_{(k+1)}\notin R|E_k\}\Pr\{y=L|E_{k+1}\}$.
- Set $f_L(x)=\Pr\{x_{(1)}\in R\}\mathbf{1}(y_{(1)}=L)+\Pr\{x_{(1)}\notin R\}\Pr\{y=L|E_1\}$.

The computation of $\Pr\{x_{(k)}\in R|E_{k-1}\}$ can be carried out by

$$\Pr\{x_{(k)}\in R|E_{k-1}\}=1-\Pr\{x_{(k)}\notin R|E_{k-1}\}=1-\frac{\binom{\ell_0^{k+1}}{n}\binom{\ell_1^{k+1}}{n}}{\binom{\ell_0^{k}}{n}\binom{\ell_1^{k}}{n}}$$

where ℓ_0^k and ℓ_1^k indicate, respectively, the numbers of $\mathcal{C}_0$ and $\mathcal{C}_1$ points in the set $\{x_{(k)},x_{(k+1)},\ldots,x_{(\ell)}\}$, with

$$\ell_0^k=\sum_{i=k}^{\ell}\mathbf{1}(y_{(i)}=0)$$

and

$$\ell_1^k=\sum_{i=k}^{\ell}\mathbf{1}(y_{(i)}=1)$$

Since for $y_{(k)}{=}0$ we have $\ell_0^{k+1}=\ell_0^k{-}1$ and $\ell_1^{k+1}=\ell_1^k$, and similarly for $y_{(k)}{=}1$, $\ell_0^{k+1}=\ell_0^k$ and $\ell_1^{k+1}=\ell_1^k{-}1$, we obtain

$$\Pr\{x_{(k)}\in R|E_{k-1}\}=\begin{cases}\dfrac{n}{\ell_0^k} & \text{if } y_{(k)}=0\\[2mm] \dfrac{n}{\ell_1^k} & \text{if } y_{(k)}=1\end{cases} \tag{12}$$

Note also that $f_0(x_i)$ and $f_1(x_i)$ can be computed simply by letting $x{=}x_i$ and using the remaining data points in the algorithm above.

The computational complexity in computing $f_0(x_i)$ and $f_1(x_i)$ can be derived by considering the complexities associated with each successive step. The computational complexity of calculating the distances $d(x_i,x_j)$ is $O(\ell^2)$. Sorting $\ell{-}1$ distances results in $O(\ell^2\log(\ell))$ complexity for each x_i, though it can be circumvented by sorting all $d(x_i,x_j)$ once and for all at the beginning albeit with the same complexity. Carrying out the computation of $\Pr\{y_i{=}0\}$ or $\Pr\{y_i{=}1\}$ is $O(\ell(\ell{-}2n))$ since it requires no more than $\ell{-}2n$ successive decompositions of the probability. The overall complexity associated with computing $f_0(x_i)$ and $f_1(x_i)$ for all x_i is therefore $O(\ell^2\log(\ell))$.

2.2. Class overlap measures based on posterior distribution estimates

Several measures of class overlap at a point $x \in \mathcal{X}$ can be computed from $f_0(x)$ and $f_1(x)$ using the expressions in Eqs. (4) and (5). As suggested by Eq. (6), $f_0(x)$ and $f_1(x)$ can be used to compute a measure $M_{\mathrm{LLR}}(x)$ that estimates the log-likelihood ratio of the two classes by

$$M_{\mathrm{LLR}}(x) = \log\left(\frac{f_0(x)}{f_1(x)}\right) \tag{13}$$

for all x with $f_0(x) \neq 0$ and $f_1(x) \neq 0$, providing the maximum likelihood classification rule when compared to 0. The overlap between the two classes is then given by the set of points in $\mathcal{X}$ where $M_{\mathrm{LLR}}(x) \simeq 0$. A major benefit of M_{LLR} is its ability to determine the specificity at which the points x_i occur within the two classes. It therefore allows dividing $\{x_i\}$ into three subsets, as those that are specific to $\mathcal{C}_0$, those that are specific to $\mathcal{C}_1$, and those that are non-specific beyond a certain specificity threshold $\alpha \lll 1$. These divisions can be obtained by assigning all the samples for which $M_{\mathrm{LLR}}(x_i) > \log((1-\alpha)/\alpha)$ to the group specific to $\mathcal{C}_0$, those for which $M_{\mathrm{LLR}}(x_i) < -\log((1-\alpha)/\alpha)$ to the group specific to $\mathcal{C}_1$, and the remaining samples to the non-specific group. The significance of α as the specificity threshold is in ensuring that no more than a fraction α of samples in either specific group is included in that group erroneously.

A second measure of class overlap can be defined in inspiration from the Henze–Penrose affinity [12,13] that computes the integral

$$\int_x \frac{2p_1(x)p_2(x)}{p_1(x)+p_2(x)}\,dx$$

for any given probability distributions $p_1(x)$ and $p_2(x)$, and goes to 1 when $p_1(x)=p_2(x)$ for all x. We define the measure $M_{\text{HP-like}}(x)$ for a sample x as a variant of the integrand above by

$$M_{\text{HP-like}}(x) = f_0(x)f_1(x) \simeq \frac{p(x|x\in\mathcal{C}_0)p(x|x\in\mathcal{C}_1)}{(p(x|x\in\mathcal{C}_0)+p(x|x\in\mathcal{C}_1))^2} \tag{14}$$

Note that over the regions of overlap, $f_0(x) \simeq f_1(x) \simeq 1/2$ and $M_{\text{HP-like}}(x)$ approaches 1/4. Conversely, for points that are highly specific to one or the other class, $M_{\text{HP-like}}(x)$ is near zero.

A final measure of overlap can be computed using the difference of $f_0(x)$ and $f_1(x)$ by

$$M_{\mathrm{Diff}}(x) = f_0(x)-f_1(x) \simeq \frac{p(x|x\in\mathcal{C}_0)-p(x|x\in\mathcal{C}_1)}{p(x|x\in\mathcal{C}_0)+p(x|x\in\mathcal{C}_1)} \tag{15}$$

Note that $M_{\mathrm{Diff}}(x)$ computes the difference of the posterior distributions of the two classes at a sample x, forming the basis of the maximum a posteriori classification rule. This measure is similar to M_{LLR} in the sense that the points of strong overlap are also given by the set of points for which $M_{\mathrm{Diff}} \simeq 0$. On the other hand, M_{Diff} can be computed for any x, even those for which $f_0(x)=0$ or $f_1(x)=0$. In addition, it can be shown that the thresholds $\pm(1-2\alpha)$ on M_{Diff} provide a division of the data into $\mathcal{C}_0-$specific, $\mathcal{C}_1-$specific, and non-specific data points for a given significance level α as well.

2.3. Selection of the optimal reference set size

The discussion above offers the class overlap measures M_{LLR}, $M_{\text{HP-like}}$, and M_{Diff} to be computed from available data using the quasi-supervised learning algorithm in Section 2.1. The accuracy at which these measures follow their true values, however, inevitably depends on the number of samples included in the reference sets from each class, denoted by n. Ideally, the best n would produce minimal class overlap or, equivalently, maximal class separation, so that M_{LLR} and M_{Diff} are never around 0, and $M_{\text{HP-like}}$ is 0 for all x_i. In addition, n should be as small as possible, for large n produces nearest neighbor classifiers that are too flexible [14,15], reducing accuracy and increasing the estimation noise. In light of these considerations, we propose the following cost functional:

$$E(n) = 4\sum_i M_{\text{HP-like}}(x_i) + 2n \tag{16}$$

to be minimized with respect to n. This cost functional represents a trade-off between a desire to penalize the choices of n that produce a large overlap between the two classes via the first term, and a competing desire to limit n to smaller values via the second term. The reasoning behind the first term follows from the main objective of the quasi-supervised learning paradigm: to identify the samples that are specific to their class of origin by minimizing the overlap. The second term incorporates the structural risk minimization principle in statistical learning that favors nearest neighbor classifiers with small reference sets for better generalization ability [14,15]. The scaling of the first term by a factor of 4 ensures that the costs incurred in the two marginal scenarios where, at the one end, $M_{\text{HP-like}}(x_i)$ $=1/4$ for all x_i indicating complete overlap, and at the other $n=\ell/2$ when the reference sets are as large as they can be (assuming $\ell_0=\ell_1$), are equal.

In order to illustrate the suitability of the cost functional in Eq. (16) in determining the best n, we have generated two random datasets, the first dataset containing 50 points drawn independently from a Gaussian distribution with zero mean and unit standard deviation, and the second dataset containing 50 points drawn independently from another Gaussian distribution with mean 1 and unit standard deviation. We have then carried out the quasi-supervised learning algorithm contrasting these two datasets for $n=1,2,3,\ldots,49$, and computed the class overlap measures M_{LLR}, $M_{\text{HP-like}}$, and M_{Diff} for all points $\{x_i\}$. By comparing the resulting class overlap measures to the true theoretical values, we have obtained the mean squared error graphs of the estimated class overlap measures for varying n. These graphs are shown in Fig. 2 along with the plot of the cost functional $E(n)$. The smallest errors on M_{LLR}, $M_{\text{HP-like}}$, and M_{Diff} were obtained at $n=3$, 3, and 2, respectively. The functional $E(n)$ attained its minimum value at $n=3$. Note that the jagged behavior observed on the mean squared error plot of M_{LLR} for large n is due to the removal of an increasing number of data points x_i for which $f_0(x_i)=0$ or $f_1(x_i)=0$ from the error computation.

2.4. Detection of abnormalities using quasi-supervised learning

The algorithm described above estimates the posterior distributions of two classes by $f_0(x_i)$ and $f_1(x_i)$ at each data vector x_i in a dataset $\{x_i,y_i\}$, $i=1,2,\ldots,\ell$, where y_i are the binary class labels. While this covers a very general family of statistical learning problems, the scenario primarily addressed in this paper refers to a case where the first class consists of samples drawn from a single reference control probability distribution, and the second containing samples drawn from the reference distribution as well as a distinct target distribution, modeling the distribution of the abnormal samples defined in some sense. In other words, from here on, we will treat the class $\mathcal{C}_0$ as representing the reference control dataset, and the class $\mathcal{C}_1$ as the unlabeled mixed dataset, and address the identification of samples in $\mathcal{C}_1$ that belong to the target distribution.

Consider the class-conditional probability distributions $p_0(x)$ and $p_1(x)$ for classes $\mathcal{C}_0$ and $\mathcal{C}_1$ given by

$$p_0(x) = p_r(x) \tag{17}$$

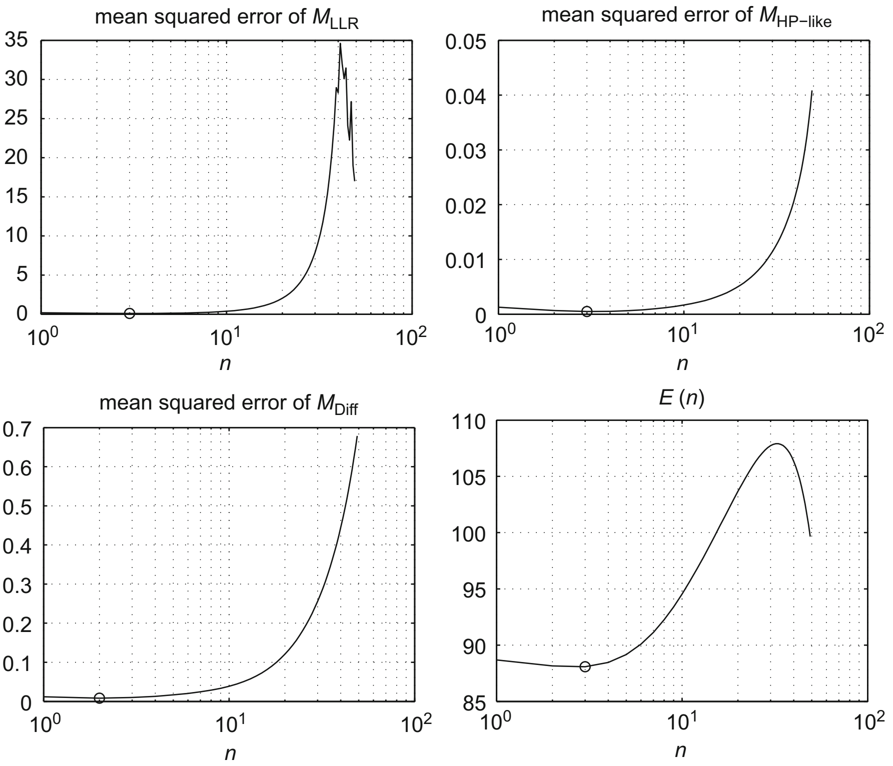

Fig. 2. The plot of $E(n)$ for the previous example for $n=1,2,\ldots,49$ along with the mean squared errors associated with the class overlap measures M_{LLR}, $M_{\text{HP-like}}$, and M_{Diff} in comparison with their true values. The smallest errors were obtained, respectively, at $n=3$, 3, and 2. The functional $E(n)$ attained its minimum value at $n=3$.

and

$$p_1(x)=(1-\lambda)p_r(x)+\lambda p_t(x) \tag{18}$$

where p_r and p_t denote the reference control and the target distributions, respectively. For $\lambda \simeq 1$, the situation converges to a conventional supervised statistical learning problem addressed by a plethora of classification methods studied extensively in the pattern recognition literature. For small λ, however, the problem of separating the two classes via classification becomes daunting due to the strong overlap between the two classes. Classification algorithms, devoting the greatest effort to sorting through the region of overlap, are prone to miss or downplay the significance of the small set of points generated from the target distribution in the mixed dataset. The quasi-supervised learning algorithm, on the other hand, directly computes estimates for the posterior distributions of the two classes at all data points. Note that under the circumstances described above, the expressions for $f_0(x)$ and $f_1(x)$ become

$$f_0(x) \simeq \frac{p_r(x)}{(2-\lambda)p_r(x)+\lambda p_t(x)} \tag{19}$$

and

$$f_1(x) \simeq \frac{(1-\lambda)p_r(x)+\lambda p_t(x)}{(2-\lambda)p_r(x)+\lambda p_t(x)} \tag{20}$$

Thus, over the points drawn from the reference distribution, the estimates f_0 and f_1 would follow each other closely around 0.5 with f_0 slightly greater than f_1, all the while maintaining $f_0(x)+f_1(x)=1$ for all x. Conversely, over the points drawn from the target distribution, f_1 would increase toward 1 at the expense of decreasing f_0, implying specificity to the second class, and hence, the target distribution.

The detection of abnormalities can be accomplished simply by thresholding the class overlap measures M_{LLR} or M_{Diff} observed over the mixed dataset with respect to a specific threshold. The samples for which these class overlap measures fall below the associated thresholds are thus labeled as abnormal. Furthermore, the detection rate can be tuned to achieve a desired false discovery rate, measured by the ratio of samples in the reference control dataset satisfying the detection criterion. Note that the samples of the reference control dataset are normal by construction, and are not subject to scrutiny by abnormality detection.

3. Results

In this section, we first present experiment results contrasting the proposed algorithm to alternative strategies from the literature on a synthetic dataset. We then demonstrate the utility of the algorithm for biomedical data analysis on real multi-color flow cytometry and multi-channel electroencephalography datasets.

3.1. Detection performance on synthetic data

In order to assess the detection performance of the proposed method, we have randomly generated reference control and mixed testing datasets of known probability distributions with a range of controlled parameters, and computed the average receiver operating characteristics curves for each parameter combination. The reference control dataset was modeled by a d-dimensional multivariate Gaussian with zero mean and identity covariance. The target distribution was also modeled as a multivariate Gaussian with identity covariance, but with mean shifted along the first dimension to 3. The reference dataset contained N samples all drawn from the reference distribution, while the mixed dataset consisted of $(1-\lambda)N$ samples drawn from the reference distribution along with λN samples drawn from the target distribution. In our experiments, we have treated the dimensionality d, the number of samples per dataset N, and the fraction of samples in the mixed dataset from the target distribution λ as control variables. In each independent repeat, we have randomly generated the reference and mixed datasets according to the parameters d, N, and λ, carried out abnormality detection using the quasi-supervised learning algorithm and computed the corresponding receiver operating characteristics curves.

In each experiment, abnormality detection via quasi-supervised learning entailed thresholding the measure M_{Diff} on the mixed dataset samples with a threshold $T \in (-1,1)$ and labeling those for which $M_{\text{Diff}}(x) \leq T$ as from the target distribution. The detection and the false alarm rates denoted by $P_{\text{D}}(T)$ and $P_{\text{FA}}(T)$ were defined as the fraction of samples from the target and reference distributions in the mixed dataset detected correctly and incorrectly. Note that the choice of the measure M_{Diff} for detection does not precondition the recognition performance of the quasi-supervised learning strategy, since an identical criterion can be devised by computing and comparing M_{LLR} against an appropriately chosen threshold as well. In both cases, the recognition is carried out in the maximum a posteriori sense, where the samples are assigned based on the posterior probabilities.

For comparison purposes, we have also trained a support vector machine classifier to separate the reference and mixed datasets [1,2]. In the classifier construction, we have used a Gaussian radial basis function kernel with

$$K(x_i,x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{21}$$

where the scale parameter σ was determined by minimizing the number of support vectors in training via a line search. In order to take into account the non-separable cases, the Lagrange multipliers of the quadratic optimization were bound from above by 100 during training, producing a soft-margin classification. The detection of the samples in the mixed dataset originating from the target distribution was carried out by thresholding the classifier underlying function

$$h(x) = \sum_i y_i \beta_i K(x,x_i) + b \tag{22}$$

by a threshold $T' \in (-\infty,\infty)$, y_i being -1 or 1 based on whether x_i belongs to the reference or mixed datasets, and β_i and b obtained by training the classifier. The mixed dataset samples for which $h(x) \geq T'$ were then identified as from the target distribution. The associated false alarm and detection rates were computed as before, respectively, as the fraction of mixed dataset samples drawn from the reference distribution detected erroneously, and the fraction of those drawn from the target distribution detected correctly.

As a second alternative, we have constructed a minimum spanning tree on all samples, and identified those of the mixed dataset that were not connected to any of the reference control samples as from the target distribution. Note that in such a minimum spanning tree, the mixed dataset samples that share edges only with other mixed dataset samples would be the ones that reduce a graph-theoretic estimate of the Henze–Penrose affinity between the datasets [12,16,13]. Note also that this strategy produces only a single false alarm rate and a single detection rate as the detection rule cannot be varied by changing a threshold as in the previous two cases, though a receiver operating characteristics curve can be constructed by joining the paired false alarm and detection rates with the origin on one side, and paired unit false alarm and detection rates on the other.

We have let $d=1, 2, 3$, $N=50, 100, 200$, and $\lambda = 0.25, 0.50, 0.75$, and carried out 20 random experiments for each combination. As the case where $\lambda = 1.00$ does not allow computation of a false alarm rate on the mixed testing dataset, it was omitted from the experiments. A representative random experiment with $d=2$, $N=50$, and $\lambda = 0.25$ is shown in Fig. 3. The reference dataset samples are marked by gray cross signs and those of the mixed dataset by gray plus signs, while the mixed dataset samples from the target distribution by dark dots. The broken line indicates the optimal Bayes detection boundary achieving 11 correct detections

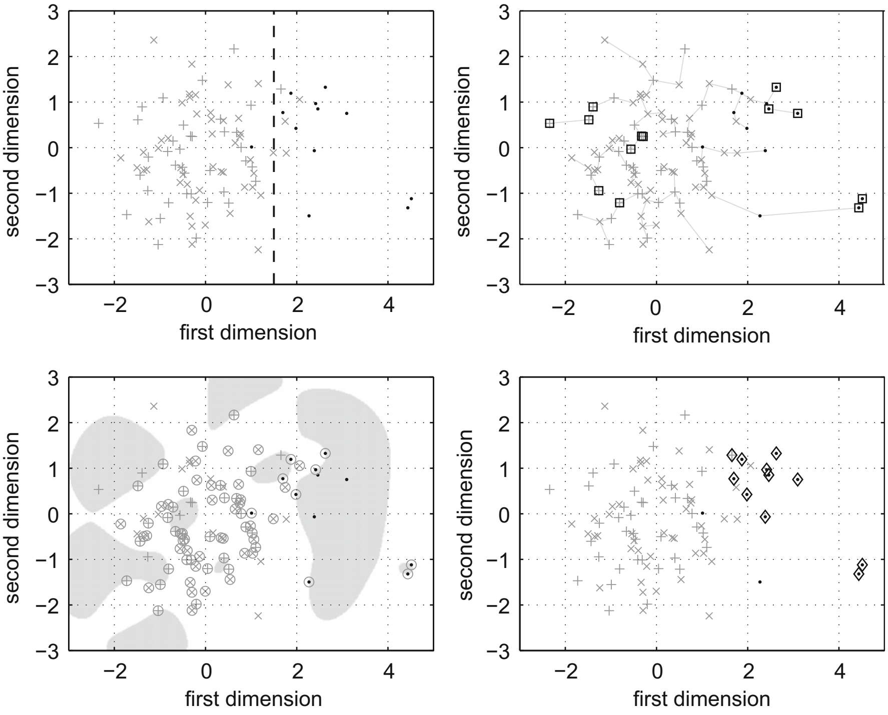

Fig. 3. Detection outcomes in a representative performance evaluation experiment with $d=2$, $N=50$, and $\lambda = 0.25$. The optimal Bayes detection is shown in (a), while the detections by the minimum spanning tree strategy, support vector machine classification, and the quasi-supervised learning algorithm are shown in (b and c).

and 1 misdetection (Fig. 3a). The minimum spanning tree strategy detects only 5 out of 12 target samples along with 8 misdetections indicated by squares (Fig. 3b). The support vector machine strategy constructs a decision boundary between the two datasets using 77 support vectors (in circles) and incurs 18 misdetections (in the shaded area) in return for 10 correct detections (Fig. 3c). The proposed quasi-supervised learning strategy incurs only one misdetection for 10 correct detections, shown by diamond signs (Fig. 3d).

In order to contrast the detection performances, we have computed the area under the average receiver operating characteristics curves of 20 independent repeats carried out for each combination of d, N, and λ. The results are shown in Table 1. For small λ, the detection performance of the support vector machine method was very poor, and it gradually improved to approach the Bayesian optimal for larger values of λ and increasing N, especially at low d. The average area under the curve values by the minimum spanning tree method generally rested well below the Bayesian optimal though they did not exhibit such strong dependence on λ. By contrast, the quasi-supervised learning method's performance closely matched the Bayesian optimal for all values of λ and N. In addition, the negative effect of the dimensionality increase on the support vector machine detection performance was not apparent in the performance of the quasi-supervised learning method.

Table 1
Performance evaluation of the quasi-supervised learning algorithm (QSL) in comparison with the graph theoretic alternative using minimum spanning trees (MST) and support vector machine classification (SVM) on synthetic data.

	MST	SVM	QSL
$N=50$			
$d=1$			
$\lambda=0.25$	0.7897	0.6531	0.9259
$\lambda=0.50$	0.8570	0.9102	0.9413
$\lambda=0.75$	0.8845	0.9607	0.9280
$d=2$			
$\lambda=0.25$	0.7625	0.6840	0.9332
$\lambda=0.50$	0.8470	0.8991	0.9505
$\lambda=0.75$	0.8638	0.9455	0.9286
$d=3$			
$\lambda=0.25$	0.7319	0.6870	0.9199
$\lambda=0.50$	0.7970	0.8342	0.9397
$\lambda=0.75$	0.8599	0.9224	0.9324
$N=100$			
$d=1$			
$\lambda=0.25$	0.7727	0.6734	0.9438
$\lambda=0.50$	0.8440	0.9648	0.9560
$\lambda=0.75$	0.8770	0.9579	0.9467
$d=2$			
$\lambda=0.25$	0.7843	0.6711	0.9323
$\lambda=0.50$	0.8200	0.8926	0.9569
$\lambda=0.75$	0.8637	0.9532	0.9485
$d=3$			
$\lambda=0.25$	0.7577	0.6302	0.9456
$\lambda=0.50$	0.8040	0.8308	0.9555
$\lambda=0.75$	0.8480	0.9277	0.9541
$N=200$			
$d=1$			
$\lambda=0.25$	0.7923	0.7700	0.9494
$\lambda=0.50$	0.8435	0.9610	0.9579
$\lambda=0.75$	0.8802	0.9633	0.9639
$d=2$			
$\lambda=0.25$	0.7740	0.6012	0.9394
$\lambda=0.50$	0.8225	0.9316	0.9624
$\lambda=0.75$	0.8677	0.9601	0.9678
$d=3$			
$\lambda=0.25$	0.7790	0.6925	0.9358
$\lambda=0.50$	0.8113	0.8589	0.9616
$\lambda=0.75$	0.8577	0.9489	0.9663

The measures represent the areas under the average receiver operating characteristics curves from 20 independent repeats. In all cases, the area under the receiver operator characteristics curve of an optimal Bayes classifier was 0.9752.

3.2. Comparative analysis of flow cytometry data

Multicolor flow cytometry is a powerful tool for rapidly and quantitatively characterizing the phenotypes of cell populations. The technology is based on fluorescently tagging each individual cell in a suspension for specific biomarkers and scanning them individually under laser illumination. This provides a high dimensional vector of features for every cell, pertaining to their size, shape, and the set of biomarkers they possess. Comparative analysis of flow cytometry data then concerns identifying different cell types in terms of these high dimensional feature vectors, or contrasting the cell distributions obtained from different individuals or at different times [17–19].

We have applied the proposed quasi-supervised learning method on a flow cytometry dataset that was originally collected and used in the study of causal relationships in human primary naïve CD4+ T cells [20]. The part of the dataset used in our experiments consists of T cells fluorescently labeled for 11 biomarkers (raf, mek1/2, Plcγ, PIP2, PIP3, Erk, akt, PKA, PKC, p38, jnk) once after a general baseline stimulation mediated by anti-CD3/CD28, and after further stimulation mediated by one of akt-inhibitor, LY294002, and Psitectorigenin. The baseline dataset contained fluorescence measurements from 853 cells, and the subsequent datasets contained fluorescence measurements from 911, 848, and 810 cells, respectively.

We have compared the akt-inhibitor, LY294002, and Psitectorigenin stimulation datasets separately to the baseline dataset after taking the feature values to the natural logarithm and computed the measures M_{LLR} and M_{Diff} at all cells. In Fig. 4, the cells with more than 97.5% specificity to their respective group are shown by dark cross signs. The histograms of the class overlap measure M_{Diff} between the baseline and post-stimulation datasets indicate that the stimulation mediated by Psitectorigenin has significantly altered the baseline cell distribution, resulting in a bi-modal histogram with most of the measurements accumulated near ± 1. The overlap with the baseline is much stronger in the post-stimulation mediated by the akt-inhibitor, as indicated by a significantly smaller number of cells specific to their own group beyond the 97.5% level. The stimulation mediated by LY294002 had no perceivable effect on the cell distributions, exposed by a complete overlap with the baseline dataset and the lack of any cells with sufficiently high group specificity. Note that while the scatter plots show the cell distributions in terms of the two features exhibiting the greatest correlation with the class overlap measure M_{Diff}, the computation of the measure itself was carried out using all 11 features simultaneously.

3.3. Pattern detection in electroencephalography data

An electroencephalography plot of an individual refers to the recordings of the electrophysiological activity of their brain measured by electrodes placed over the scalp [21]. The small voltages measured at each electrode act as channel readings. From a statistical learning point of view, the collection of readings from many channels acquired at a specific time instant organized into a column vector corresponds to a data sample reflecting the individual's brain activity pattern at that instant. These samples can then be analyzed in a variety of clinical settings using pattern recognition methods.

We have used the quasi-supervised learning method to contrast the brain activity patterns of an alcoholic and a healthy control subject on a previously published electroencephalography dataset [22]. The data consist of 64-channel 256 Hz electroencephalography recordings of 1 s duration from 10 trials, under visual stimuli where the subjects were presented by either a matching or non-matching pair of images. We have contrasted the matching dataset samples to the non-matching dataset samples separately for each subject and computed the M_{Diff} measure for each sample after normalization to have zero mean and unit standard deviation in all 64 channels to remove any potential magnitude bias. In both cases, the datasets contained 2560 samples of 64 dimensions each.

The histograms of the class overlap measure M_{Diff} shown in Fig. 5 reveal the differences in the alcoholic subject's responses to matched and non-matched visual stimuli compared to those of

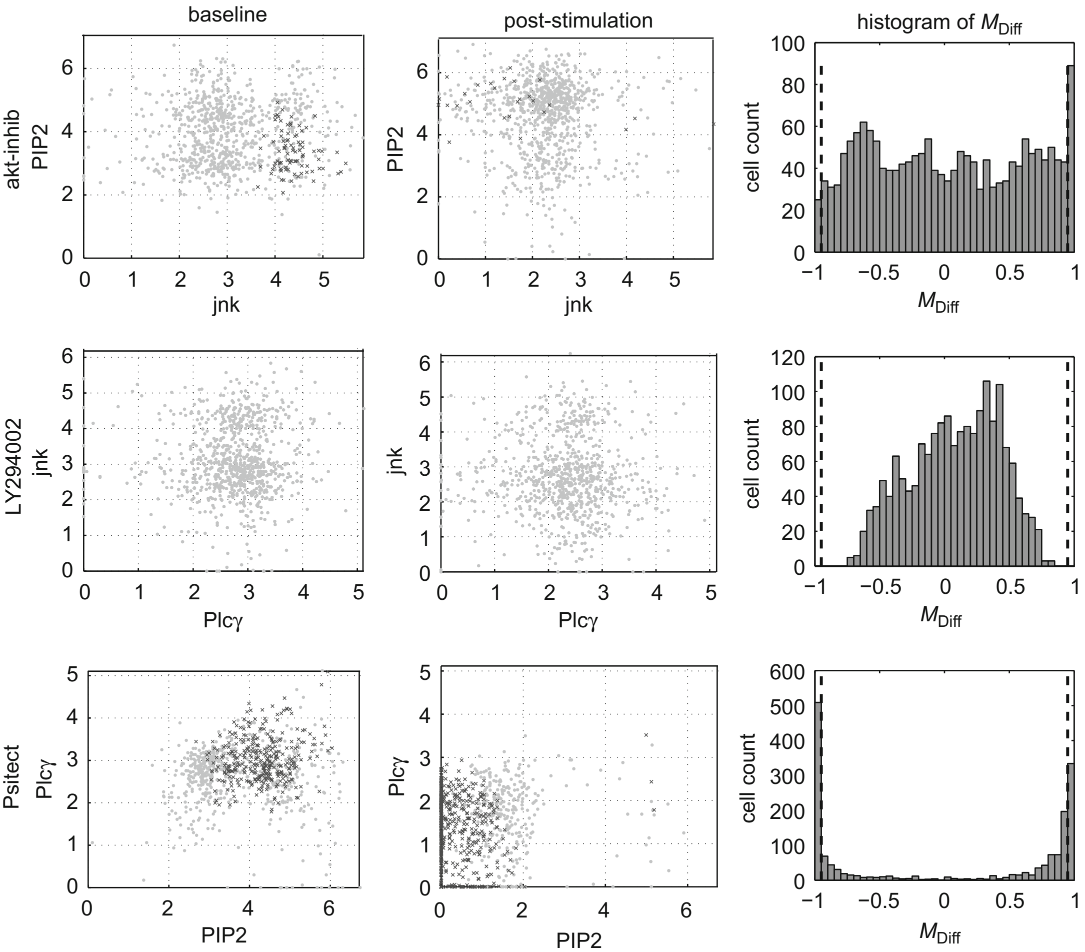

Fig. 4. Results of the flow cytometry experiments. The cells with specificity beyond 97.5% are shown darker than the others. In the histograms, the 97.5% specificity thresholds are indicated by vertical broken lines. The scatter plots are shown with respect to the parameters exhibiting the greatest correlation with specificity measure M_{Diff}.

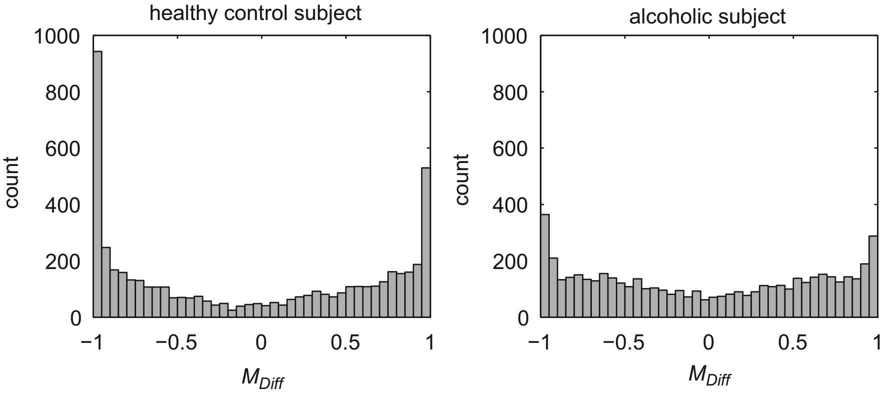

Fig. 5. Histograms of the class overlap measures M_{Diff} comparing the electroencephalography activity in matched and non-matched visual stimuli in a healthy control and an alcoholic subject. The large number of M_{Diff} observations near ± 1 attest the significant difference in the brain activity patterns in the healthy control subject in response to the difference in visual stimulus (left). A more diffuse histogram of M_{Diff} suggests a reduction in the stimulus-specific brain activity in the alcoholic subject (right).

the control subject. The brain activity patterns of the healthy control subject show a clear distinction between the two stimulus types, as indicated by the number of class overlap measures near ± 1. Conversely, the brain activity patterns of the alcoholic subject during matched and non-matched stimuli overlap to a larger extent as evidenced by a more diffuse histogram. These results are consistent with the evidence in the literature documenting significant differences in brain activity in response to various stimuli in alcoholism [23].

4. Conclusion

We have presented a novel statistical learning algorithm that contrasts two datasets by computing estimates for the posterior probability of their samples of belonging in either dataset. When one of the dataset contains samples of one class only and the other an unlabeled mixture, the algorithm instantiates a quasi-supervised learning paradigm whereby the samples of a second class in the mixed dataset are identified automatically based on the computed posterior probabilities. In performance evaluation experiments on synthetic target detection data, the proposed method outperformed alternative strategies based on support vector machine classification and minimum spanning trees for varying dataset size, overlap, and dimensionality.

The proposed paradigm is ideally suited to a dichotomous biomedical data analysis setting of abnormal versus normal since it does not require exact knowledge of which samples are abnormal in the mixed dataset. Given the difficulties and drawbacks of collecting manually labeled ground truth datasets for supervised learning in biomedical applications, quasi-supervised learning proves to be a viable alternative with minimal manual input. These strengths were demonstrated by experiments on automated comparison of cell distributions in multi-color flow cytometry data and brain activity pattern analysis on 64-channel electroencephalography data.

In a wider perspective, estimation of posterior probabilities from available data makes the proposed algorithm suitable also for general statistical learning tasks such as classification as a model-free alternative to existing techniques. Especially in cases where the data do not allow perfect separation of the different classes, the algorithm can be expected to outperform off-the-shelf classification algorithms as it will avoid searching for a separation boundary optimized according to some criterion. In extreme cases, the estimated posterior probabilities can be followed by regression algorithms to reduce the estimation noise. Extension of the algorithm from binary to multi-class classification is also straightforward, with minor modifications on the conditional probability decompositions to take into account the presence of additional class labels.

As is the case with all statistical learning methods, however, the proposed scheme is not immune against the well-established issues with feature selection and data normalization. Indeed, it is conceptually possible to undertake feature selection iteratively within the algorithm by gradually removing the features that are ineffectual towards the estimated posterior probabilities. On the other hand, the specifics of such a strategy would inevitably be application-dependent, and must be addressed separately in each case.

References

[1] C. Cortes, V.N. Vapnik, Support vector networks, Machine Learning 20 (1–2) (1995) 273–297.

[2] V.N. Vapnik, Statistical Learning Theory, Wiley, 1998.

[3] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, Information Theory 13 (1) (1967) 21–27.

[4] K. Fukunaga, L.D. Hostetler, k-nearest-neighbor bayes risk estimation, IEEE Transactions on Information Theory 21 (3) (1975) 285–293.

[5] S. Haykin, Neural Networks and Learning Machines, third ed., Prentice-Hall, 2008.

[6] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition. Wiley Series in Probability and Statistics, Wiley-Interscience, 2004.

[7] K.C. Gowdaa, G. Krishnab, Learning with a mutualistic teacher, Pattern Recognition 11 (5–6) (1979) 383–390.

[8] J.A. Flanagan, Context awareness in a mobile device: ontologies versus unsupervised/supervised learning, in: Proceedings of AKKR05, 2005, pp. 167–170.

[9] O. Chapelle, B. Schölkopf, A. Zien, Introduction to Semi-Supervised Learning, The MIT Press, 2006.

[10] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley-Interscience, 2000.

[11] B. Karaçalı, A. Tözeren, Automated detection of regions of interest for tissue microarray experiments: an image texture analysis, BMC Medical Imaging 7 (2) (2007).

[12] N. Henze, M. Penrose, On the multivariate runs test, in: Annals of Statistics, 1999, pp. 290–298.

[13] H. Neemuchwala, A.O. Hero, Entropic graphs for registration, in: R.S. Blum, Z. Liu (Eds.), Multi-Sensor Image Fusion and its Applications, Marcel Dekker, Inc., 2005, pp. 185–235.

[14] B. Karaçalı, H. Krim, Fast minimization of structural risk by nearest neighbor method, IEEE Transactions on Neural Networks 14 (1) (2003) 127–137.

[15] B. Karaçalı, R. Ramanath, W. Snyder, Structural risk minimization-based nearest neighbor classifier, Pattern Recognition Letters 25 (1) (2004) 63–71.

[16] A.O. Hero, B. Ma, O.J. Michel, J. Gorman, Applications of entropic spanning graphs, IEEE Signal Processing Magazine 19 (5) (2002) 85–95.

[17] S.P. Perfetto, P.K. Chattopadhyay, M. Roederer, Seventeen-colour flow cytometry: unravelling the immune system, Nature Reviews Immunology 4 (8) (2004) 648–655.

[18] V. Gattei, M. Degan, S. Russo, R. Bomben, M.D. Bo, M. Rupolo, F. Buccisano, G.D. Poeta, P. Sonego, A. Zucchetto, Immunophenotypic clustering of b-cell chronic lymphocytic leukemia (b-cll) reveals a good prognosis disease subset characterized by the coordinated over-expression of cd62l, cd54, cd49c, cd25 and cd55, Journal of Clinical Oncology 22 (14S) (2004) 6567.

[19] M. Roederer, A. Treister, W. Moore, L. Herzenberg, Probability binning comparison: a metric for quantitating univariate distribution differences, Cytometry 45 (1) (2001) 37–46.

[20] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, G.P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data, Science 308 (5721) (2005) 523–529.

[21] C.D. Binnie, P.F. Prior, Electroencephalography, Journal of Neurology, Neurosurgery, and Psychiatry 57 (11) (1994) 1308–1319.

[22] L. Ingber, Statistical mechanics of neocortical interactions: canonical momenta indicators of electroencephalography, Physical Review E 55 (4) (1997) 4578–4593.

[23] B.H. PorjeszB, Alcoholism and human electrophysiology, Alcohol Research & Health 27 (2) (2003) 153–160.

Bilge Karaçali has received his BS on Electrical and Electronics Engineering from Bilkent University, and MS and PhD on Electrical Engineering from the North Carolina State University in 1999 and 2002. He is currently with the Electrical and Electronics Engineering Department at İzmir Institute of Technology as an associate professor.